Statistical Analysis in Downscaling Climate Models:

Wavelet and Bayesian Methods in Multimodel Ensembles

by

Yihua Cai, M.A.

Report

Presented to the Faculty of the Graduate School
of the University of Texas at Austin
in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science in Statistics

The University of Texas at Austin
August 2009

The Report committee for Yihua Cai
Certifies that this is the approved version of the following report:

Statistical Analysis in Downscaling Climate Models: Wavelet and Bayesian Methods in Multimodel Ensembles

APPROVED BY
SUPERVISING COMMITTEE:

Supervisor: Paul Damien

Robert E. McCulloch

Acknowledgements

My utmost gratitude goes to my report advisor, Dr. Paul Damien and Dr. Robert E. McCulloch, for their expertise, kindness, and most of all, for their support. I believe that one of my main gains in this 2-years program was working with Dr. Damien and Dr. McCulloch, and starting my research. I would like to express my appreciation to my graduate advisor, Dr. Daniel A. Powers who leaded me into Statistics, the area I am working in.

Statistical Analysis in Downscaling Climate Models: Wavelet and Bayesian Methods in Multimodel Ensembles

by

Yihua Cai, M.S.Stat.

The Univeristy of Texas at Austin, 2009

SUPERVISOR: Paul Damien

Various climate models have been developed to analyze and predict climate change; however, model uncertainties cannot be easily overcome. A statistical approach has been presented in this paper to calculate the distributions of future climate change based on an ensemble of the Weather Research and Forecasting (WRF) models. Wavelet analysis has been adopted to de-noise the WRF model output. Using the de-noised model output, we carry out Bayesian analysis to decrease uncertainties in model CAM_KF, RRTM_KF and RRTM_GRELL for each downscaling region.

CONTENTS

Statistical Analysis in Downscaling Climate Models: Wavelet and Bayesian Methods in Multimodel Ensembles

Introduction

People have seen a growing discussion of climate change and global warming in recent years. Various climate models have been developed to analyze and predict climate change. However, model uncertainties cannot be easily overcome. In its Synthesis Report, the Intergovernmental Panel on Climate Change (IPCC)[1] emphasizes uncertainties in climate change research.

Uncertainty itself has various levels. The uncertainty in the report is defined as that if reduced, may lead to new and robust findings (IPCC 2001). There are lots of reasons why the climate modeling community is left behind. Uncertainties could stem from "epistemic" or "stochastic" sources (Dessai, S. and Hulme, M. 2004). Epistemic sources of uncertainty are those that can be reduced by further study, improving our state of knowledge, etc. Stochastic sources of uncertainty are items such as variability in the system, the chaotic nature of the climate system, and the indeterminacy of human systems, which can hardly be reduced or simulated.

Climate models differ considerably in their estimates of the strength of different feedbacks in the climate system. In fact, many of the key uncertainties are concerned with the quantification of the magnitude and/or timing of the response. Also, the confidence in projections is higher for some variables (e.g. temperature) than for others (e.g. precipitation) (IPCC, AR4). Since each model has different variables, different confidence level for each variable would magnify the uncertainty in climate modeling. Other factors, such as climate data coverage remains limited in some regions and there is a notable lack of geographic balance in data and literature on observed changes in natural and managed systems bring more uncertainties into climate research.

There is a strong need for studying aspects of the uncertainty of climate projections. New techniques have been developed for quantifying uncertainty in climate model projections (Tebaldi et al. 2005). The emphasis of this paper is given to quantifying regional uncertainty. Regional projections of impacts are most needed by decision-makers, and yet are not easily extracted from global climate model simulations.

There is interest in climate research in moving from single-value predictions to probability forecasts. In fact, presenting a probability distribution of future climate is a more flexible approach for drawing inferences and facilitating decision making (Tebaldi et al. 2005). This paper adopts wavelet analysis and Bayesian inference in studying model uncertainties. Bayesian method is a natural way to do probability inference, because of its simplicity and efficiency. Wavelet analysis de-noises the input data, and helps the results converge better.

---

[1]. The panel was established in 1988 by the World Meteorological Organization (WMO) and the United Nations Environment Programme (UNEP), two organizations of the United Nations.

Reliability Ensemble Average Method

REA is an efficient statistical strategy that has been adopted by researchers before Bayesian method has been proposed. It has been recommended that a small set of complementary difference measures can represent an objective and meaningful description of a model's ability to reproduce reliable observations precisely or accurately. The core of this set of difference measures is made up of the root-mean-square error. Giorgi and Mearns's reliability ensemble average method takes into account two "reliability criteria": the performance of the model in reproducing present-day climate ("model performance" criterion) and the convergence of the simulated changes across models ("model convergence" criterion) (Giorgi and Mearns 2002). In the REA method, the weighted average change, $\Delta\tilde{T} = \tilde{A}(\Delta T) = \frac{\sum_i R_i \Delta T_i}{\sum_i R_i}$ (2.1)

$\tilde{A}$ denotes the REA averaging.

R_i is a function of $R_{B,i}$ and $R_{D,i}$.

$$R_i = [(R_{B,i})^m \times (R_{D,i})^n]^{[1/(m\times n)]} = \left\{ [\frac{\varepsilon_T}{abs(B_{T,i})}]^m [\frac{\varepsilon_T}{abs(D_{T,i})}]^n \right\}^{[1/(m\times n)]} \quad (2.2)$$

$B_{T,i}$ is the bias in simulating present-day temperature, which is defined as the difference between simulated and observed mean temperature for the present-day.

$R_{D,i}$ is a factor that measures the model reliability in terms of the distance ($D_{T,i}$) of the change calculated by a given model from the REA average change, that is the higher the distance, the lower the model reliability (Giorgi and Mearns 2002).

Applying an iterative procedure in order to obtain the final weighted mean, they adopt a statistical strategy which is known as iteratively reweighted least squares (Green 1984). Tebaldi et al. claimed that Bayesian analysis of model ensemble could present a probability distribution of future climate for drawing inference, which is more efficient and comprehensive (Tebaldi et al. 2005).

Bayesian Approach

Bayes' theorem, an important theorem of probability theory is widely used in statistical inference. Suppose that there is a probability model $f(x \mid \theta)$ for data x, and also suppose we summarize our beliefs about θ in a prior density $\pi(\theta)$, the distribution of θ before we have the observations. This implies that we think of the unknown value θ that underlies our data as the outcome of a random variable whose density is $\pi(\theta)$, just as our probability model is that the data x are the observed value of a random variable *X* with density $f(x \mid \theta)$. Once the data have been observed, our beliefs about θ are contained in its conditional density given that *X* = x.

$$\pi(\theta \mid x) = \frac{f(x \mid \theta)\pi(\theta)}{f(x)} \quad (3.1)$$

where $f(x) = \int f(x \mid \theta)\pi(\theta)d\theta$ (3.2)

$f(x \mid \theta)$ is the likelihood for θ based on *x*, so that, in terms of θ, we have posterior $\propto$ prior $\times$ likelihood, that is,

$$\pi(\theta \mid x) \propto \pi(\theta) \cdot f(x \mid \theta) \quad (3.3)$$

The parameter θ is an index of the family of possible distribution for the data. For classical statisticians, the parameters are fixed. However, Bayesians treat parameters as random variables. The Bayesians incorporate the information about the parameter into the analysis through a density $\pi(\theta)$.

In some cases, the posterior distributions are not members of any known distribution families. Markov Chain Monte Carlo (MCMC) simulation is used to generate a large number of sample values. By simulating from the highly dimensional distribution of the unknown quantities, MCMC helps us approximate the parameters and statistics we are interested in. Markov chain is a stochastic process with the Markov property, which means that future states depend only on the present state, and are independent of past states. In other words, Markov chains are processes describing trajectories where successive quantities are described probabilistically according to the value of their immediate predecessors. More than that, these processes tend to equilibrium and the limiting quantities follow an invariant distribution. MCMC techniques enable simulation from a distribution by embedded it as a limiting distribution of a Markov chain and simulating from the chain until it approaches equilibrium (Gamerman and Lopes 2006).

Gibbs sampling is based only on elementary properties of Markov chains. The Gibbs sampler is a technique for generating random variables from a (marginal) distribution indirectly, without having to calculate the density.

Gibbs sampling can be described in the following way (Casella, G. and George, E. 1992):

Starting with a pair of random variables (X, Y), the Gibbs sampler generates a sample from $f(x)$ by sampling instead from the conditional distributions $f(x \mid y)$ and $f(y \mid x)$. This is done by generating a "Gibbs sequence" of random variables

1. Initialize the iteration counter of the chain $j = 1$ and set initial values $\theta^{(0)} = (\theta_1^{(0)}, ..., \theta_m^{(0)})'$;

2. Obtain new value $\theta^{(j)} = (\theta_1^{(j)}, ..., \theta_m^{(j)})'$ from $\theta^{(j-1)}$ through successive generation of values

$$\theta_1^{(j)} \sim f(\theta_1 \mid \theta_2^{(j-1)}, ..., \theta_m^{(j-1)}) \qquad (3.4)$$

$$\theta_2^{(j)} \sim f(\theta_2 \mid \theta_1^{(j)}, \theta_3^{(j-1)}, ..., \theta_m^{(j-1)}) \qquad (3.5)$$

$$\theta_m^{(j)} \sim f(\theta_m \mid \theta_1^{(j)}, ..., \theta_{m-1}^{(j)}) \qquad (3.6)$$

3. Change counter j to $j+1$ and return to step 2 until convergence is reached.

To obtain an approximate sample from $f(x)$, we can generate *n* independent Gibbs sequences of length *k*. Using the final value of $X_k^{'}$ from each sequence, we have the an approximate iid sample from $f(x)$ (Gelfand and Smith 1990).

Different weather and climate research models give different predictions about the present and future temperature. Some of them are more accurate than others. In this article, we are interested in eliminating model uncertainties, which would help improve model accuracy.

Using Bayesian analysis, model uncertainty becomes our parameter. X_i is the present temperature from weather research models; Y_i is the projected future temperature from weather research models.

We assume Gaussian distribution for X_i, Y_i,

$$X_i \sim N(\mu, \sigma_i^{-1}) \qquad (3.7)$$

$$Y_i \sim N(\nu, \lambda_i^{-1}) \qquad (3.8)$$

Both X_i and Y_i have a Gaussian distribution. In (3.7) and (3.8), μ and ν are the mean values of present and future temperatures in a specific region. Focusing on eliminating model uncertainties, we are more interested in the posterior distributions of σ_i and λ_i, because σ_i^{-1} and λ_i^{-1} are the variances of X_i and Y_i. We assume σ_i and λ_i have prior distributions $Gamma(\alpha, \beta)$. The likelihood function comes from the model output, which are the projected temperatures.

The models used in this study is the Weather Research and Forecasting (WRF) Model with Advanced Research WRF (ARW) dynamic core version 2.2. WRF is a next-generation, limited-area, non-hydrostatic, with terrain following eta-coordinate mesoscale modeling system designed to serve both operational forecasting and atmospheric research needs. The main physical options we used include the new Kain-Fritsch (KF) convective parameterization; Dudhia shortwave radiation and Rapid Radiative Transfer Model (RRTM) longwave radiation; the Community Atmosphere Model (CAM), which is the latest in a series of global atmosphere models developed at the National Center for Atmospheric Research (NCAR) for the weather and climate research communities; Grell scheme, which is based on the rate of destabilization or quasi equilibrium. Combinations of different schemes are studied: CAM-KF, RRTM-KF and RRTM-GRELL. Each combination could be treated as a single model. The United States of America has been divided into seven regions for research purpose: northwest, central, northeast, midwest, southwest, texas and southeast.

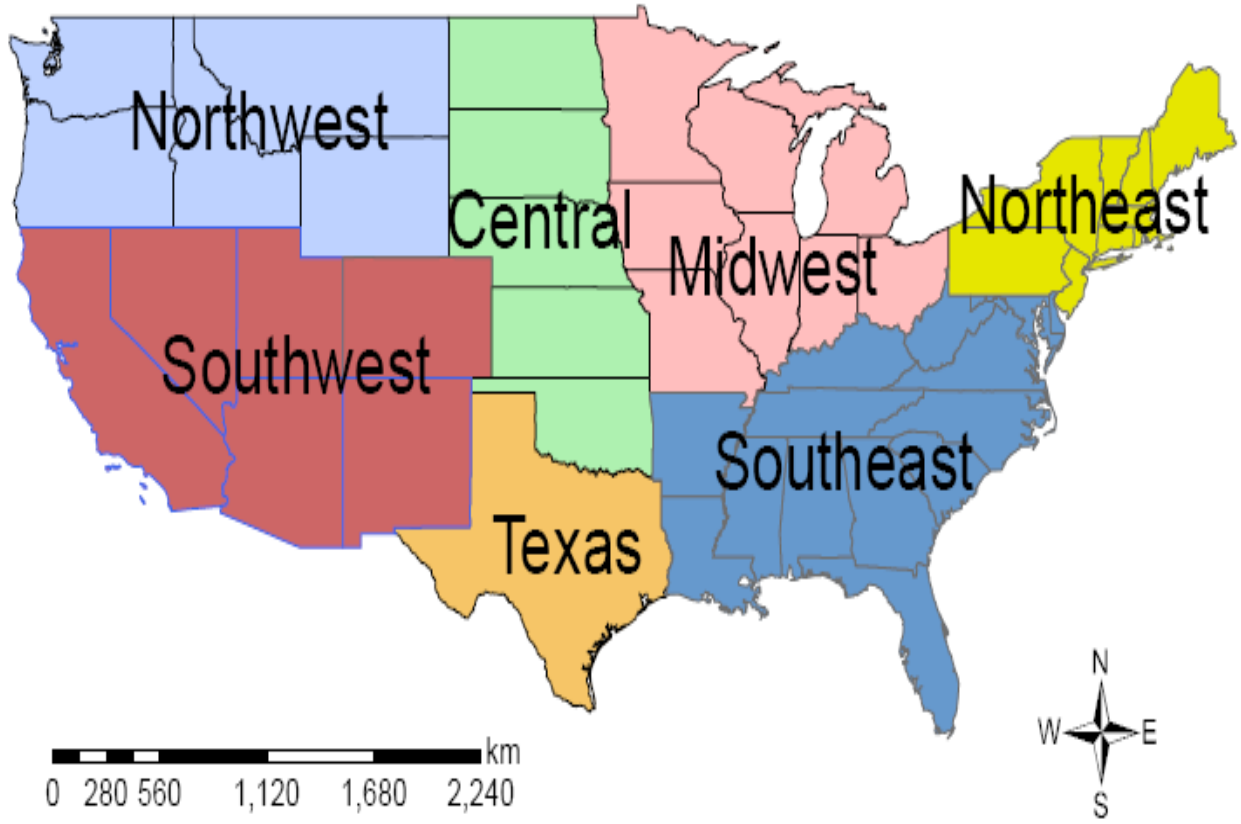

Fig 1. Regions of WRF Model Output.

The parameters μ and ν, the mean values of present and future temperature have uniform distribution. Even if these priors are improper, the form of the likelihood model ensures that the posterior is a proper density function. Alternative analyses in which the prior distribution is restricted to a finite, but sufficiently large, interval, do not in practice produce different results (Tebaldi et al., 2005).

The σ_i and λ_i ($i = 1, 2, 3$) have Gamma distribution $Gamma(\alpha, \beta)$. That is the prior distribution of σ_i and λ_i.

$$\pi(\sigma) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \sigma^{\alpha-1} e^{-\beta\sigma} \quad (3.9)$$

$$\pi(\lambda) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} \quad (3.10)$$

To make it simple, we choose $\alpha = \beta = 0.001$.

The prior and posterior distributions belong to the same class of distributions. The distribution of the parameters can be updated without any complicated calculation. The preservation of the distribution after updating in the same class defines conjugacy.

For Gamma distribution $f(x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$ (3.11)

$$E(X) = \frac{\alpha}{\beta}, \quad (3.12)$$

$$Var(X) = \frac{\alpha}{\beta^2}. \quad (3.13)$$

$$E(\sigma) = E(\lambda) = \frac{\alpha}{\beta} = \frac{0.001}{0.001} = 1, \quad (3.14)$$

and $Var(\sigma) = Var(\lambda) = \frac{\alpha}{\beta^2} = \frac{0.001}{0.001^2} = 1000$. (3.15)

Large variance means non informative quality.

$$\pi(\sigma \mid x) = \frac{f(x \mid \sigma) f(\sigma)}{f(x)} = \frac{f(x \mid \sigma) f(\sigma)}{\int f(x \mid \sigma) f(\sigma) d\sigma} \quad (3.16)$$

Since $X_i \sim N(\mu, \sigma_i^{-1})$, $f(X_i \mid \sigma_i) = \frac{\sqrt{\sigma_i}}{\sqrt{2\pi}} \exp(-\frac{(x_j - \mu)^2 \sigma_i}{2})$, (3.17)

and $f(\sigma_i) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \sigma_i^{\alpha-1} e^{-\beta\sigma_i}$. (3.18)

$$\pi(\sigma_i \mid X_i) = \frac{\prod_{j=1}^{n} \frac{\sqrt{\sigma_i}}{\sqrt{2\pi}} \exp(-\frac{(x_j - \mu)^2 \sigma_i}{2}) \frac{\beta^{\alpha}}{\Gamma(\alpha)} \sigma_i^{\alpha-1} e^{-\beta\sigma_i}}{\int \prod_{j=1}^{n} \frac{\sqrt{\sigma_i}}{\sqrt{2\pi}} \exp(-\frac{(x_j - \mu)^2 \sigma_i}{2}) \frac{\beta^{\alpha}}{\Gamma(\alpha)} \sigma_i^{\alpha-1} e^{-\beta\sigma_i}} \quad (3.19)$$

$$\sigma_i \sim Gamma((n\alpha - \frac{n}{2} + 1), \ (n\beta + \frac{\sum (x_j - \mu)^2}{2})) \quad (3.20)$$

$$E(\sigma_i \mid X_i) \approx \frac{n\alpha - \frac{n}{2} + 1}{n\beta + \frac{\sum (x_j - \mu)^2}{2}}. \quad (3.21)$$

(3.7) through (3.21) show how to get the posterior distribution of σ_i by Bayes's theorem directly. Also, (3.20), the posterior distribution of σ_i is from Gamma family, which is easy for us to calculate. In this case, no sophisticated approach is needed to calculate the posterior distribution.

However, Tebaldi et al. adopted a different strategy, which is more comprehensive. We would need Gibbs sampler to carry out the calculation. Assuming there is relation between variance of current temperature and future temperature, we would have:

$$X_i \sim N(\mu, \sigma_i^{-1}) \quad (3.22)$$

$$Y_i \sim N(\nu, (\theta\lambda_i)^{-1}) \quad (3.23)$$

Applying Bayes's theorem to the likelihood and priors, we would have the joint distribution of $\sigma, \lambda, \theta, \mu, \nu$.

$$f(\sigma, \lambda, \theta, \mu, \nu) \propto \prod_{i=1}^{3} \lambda_i^{\alpha-1} e^{-b\lambda_i} \lambda_i \theta^{\frac{1}{2}} \exp\{-\frac{\lambda_i}{2}[(x_i - \mu)^2 + \theta(y_i - \nu)^2]\} \theta^{c-1} e^{-d\theta} \exp\{-\frac{\lambda_0}{2}(x_0 - \mu)^2\}$$

(3.24).

MCMC simulation is used to generate a large number of sample values for all parameters from (3.24).

Fixing some groups of parameters and considering the conditional posterior for the others, we can find out how the posterior distribution synthesizes the data and the prior assumption (Tebaldi et al. 2005).

$$\hat{\mu} = (\sum_{i=0}^{3} \lambda_i x_i) / (\sum_{i=0}^{3} \lambda_i) \quad (3.25)$$

and variance

$$(\sum_{i=0}^{3} \lambda_i)^{-1} \quad (3.26)$$

The conditional distribution of ν, fixing all other parameters is a Gaussian distribution with mean

$$\hat{\nu} = (\sum_{i=0}^{3} \lambda_i Y_i) / (\sum_{i=0}^{3} \lambda_i) \quad (3.27)$$

and variance

$$(\theta \sum_{i=0}^{3} \lambda_i)^{-1} . \quad (3.28)$$

An approximation to the mean of the posterior distribution of the λ_i s is

$$E(\lambda_i \mid \{x_0,..., x_3, y_1,..., y_3\}) \approx \frac{\alpha + 1}{\beta + \frac{1}{2}[(x_i - \hat{\mu})^2 + \theta(y_i - \hat{\nu})^2]} . \quad (3.29)$$

$| X_i - \hat{\mu} |$ and $| Y_i - \hat{\nu} |$ correspond to the *bias* and *convergence* criteria. $| Y_i - \hat{\nu} |$ measures the distance of the *i*th model future response form the overall average response. The similar scenario has been considered in Giorgi and Mearns (2002). The important difference from REA for the Bayesian model is that the distance is based on the future projection (Y_i) rather than the temperature change ($Y_i - X_i$). As for the bias term, notice that in the limit, if we let $\lambda_0 \to \infty$, $\hat{\mu} \to X_0$, and the bias term becomes in the limit $| X_i - X_0 |$, the same definition of bias as in the REA analysis.

Assuming there is correlation between present and future temperature, Tebaldi et al. claimed that the future and present temperature is linked by a linear regression equation, and that is equivalent to assuming that (X_i, Y_i) are jointly normal. The assumption for X_i and Y_i:

$$X_i = \mu + \eta_i \quad (3.30)$$

Where $\eta_i \sim N(0, \lambda_i^{-1})$; (3.31)

$$Y_i = \nu + \beta_x (X_i - \mu) + \xi_i / \sqrt{\theta} \quad (3.32)$$

Where $\xi_i \sim N(0, \lambda_i^{-1})$. (3.33)

β_x introduces a direct or inverse relation between $X_i - \mu$ and $Y_i - \nu$. A value of β_x equal to one translates into conditional independence of these two quantities, while values greater than or less than one

would imply positive or negative correlation between them. Given the relation between X_i and Y_i, the posterior mean λ_i approximately equals:

$$\frac{\alpha+1}{\beta+\frac{1}{2}\{(x_i-\hat{\mu})^2+\theta[y_i-\hat{\nu}-\beta_x(x_i-\hat{\mu})]^2\}}. \quad (3.34)$$

Table 1. The model output of present temperature for different regions.

	CAM_KF	RRTM_KF	RRTM_GRELL
Southwest	12.0322	11.8361	11.7551
Northwest	6.6609	6.7214	6.5326
Texas	17.9983	18.1092	17.9019
Central	12.0323	12.4033	12.2348
Midwest	9.131	9.6605	9.6123
Southeast	16.6465	16.9806	16.951
Northeast	7.1099	7.5812	7.5237

Table 2. The model output of future temperature (2050) for different regions.

	CAM_KF	RRTM_KF	RRTM_GRELL
Southwest	13.7088	13.4591	13.3551
Northwest	8.1701	8.2094	8.0078
Texas	20.0633	20.1692	19.8946
Central	13.9685	14.3092	14.1242
Midwest	12.0511	12.4641	12.347
Southeast	18.5607	18.9577	18.8543
Northeast	9.6493	10.096	9.9813

The Bayesian analysis yields posterior distribution of the uncertain quantities. The posterior distributions for regional temperature change and for a suite of other parameters provide us sufficient information. Adopting Bayesian approach, Tebaldi et al. showed their model has less uncertainty and converges better than REA model advocated by Giorgi and Mearns. Further analysis reveals that output of the Bayesian approach developed by Tebaldi et al. can be improved if wavelet analysis is used.

De-Noising Method

The mean square error of an estimator $\hat{\theta}$ defines as $MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$ (4.4). A little bit calculation helps us find that $MSE(\hat{\theta}) = Var(\hat{\theta}) + (Bias(\hat{\theta}))^2$ (4.5). Many statistical techniques simply optimize the mean-squared error, which demands a tradeoff between bias and variance. Mean squared error is sum of bias and variance. Given the value of mean squared error, we cannot decrease bias without increasing variance. These optimizations exhibit noise-induced structures, which gives rise to interpretational difficulties.

Donoho and Johnstone proved that their estimator has an optimality property with respect to mean squared error for estimating functions of unknown smoothness at a point (Donoho and Johnstone 1992). In the wavelet analysis, De-noising describes various schemes which reject noise by damping or thresholding in the wavelet domain. A formal interpretation of the term "De-Noising" has been proposed by David Donoho (Donoho 1995).

$$d_i = f(t_i) + \sigma \cdot z_i \quad i = 0,...,n-1 \quad (4.1)$$

$$t_i = i/n \quad (4.2)$$

z_i iid $N(0, 1)$ Gaussian white noise, and σ is a noise level.

Donoho's interpretation of "De-Noising" is that to optimize the mean-squared error.

$$n^{-1}E(\hat{f} - f)^2 = n^{-1}\sum_{i=0}^{n-1} E(\hat{f}(i/n) - f(i/n))^2 . \quad (4.3)$$

With high probability, $\hat{f}_n^*$ is at least as smooth as f, with smoothness measured by any of a wide range of smoothness measures. $\hat{f}_n^*$ achieves almost the minimax mean square error over every one of a wide range of smoothness classes, including many classes where traditional linear estimators do not achieve the minimax rate. (Donoho 1995).

The one-dimensional model: $s(n) = f(n) + \sigma \cdot e(n)$, where time n is equally spaced. $e(n)$ is a Gaussian white noise $N(0,1)$ and the noise level is supposed to be equal to one. From a statistical point of view, this model is a regression model over time and the model can be viewed as a nonparametric estimation of the function f using orthogonal basis.

The general de-noising procedure involves three steps: decompose, threshold detail coefficients and reconstruct.

1. decompose: choose a wavelet, choose a level N. Compute the wavelet decomposition of the signals s at level N.

2. Threshold detail coefficients: For each level form 1 to *N*, select a threshold and apply soft thresholding to the detail coefficients.

3. Reconstruct: Compute wavelet reconstruction using the original approximation coefficients of level *N* and the modified detail coefficients of levels from one to *N*.

Matlab, a software package has been used to do wavelet analysis. The function $yt = wthresh(y, sorh, thr)$ can be used to do thresholding. Depending on the $sorh$ option, this function returns soft or hard thresholding of input y. Hard thresholding, which is the simplest method can be described as the usual process of setting to zero the elements whose absolute values are lower than the threshold. Soft thresholding, which has nice mathematical propterties is an extension of hard thresholding, first setting to zero the elements whose absolute values are lower than the threshold, and then shrinking the nonzero coefficients towards zero.

Fig 2. Output of model CAM-KF, central U.S.

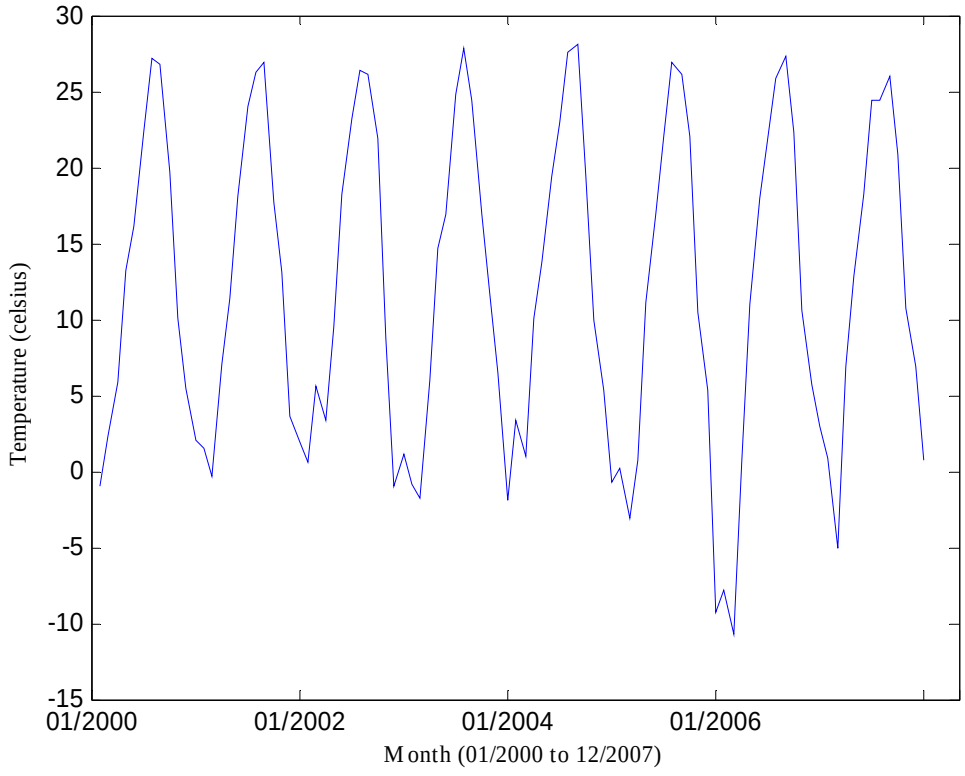

Fig 3. Output of model CAM-KF, central U.S. being de-noised.

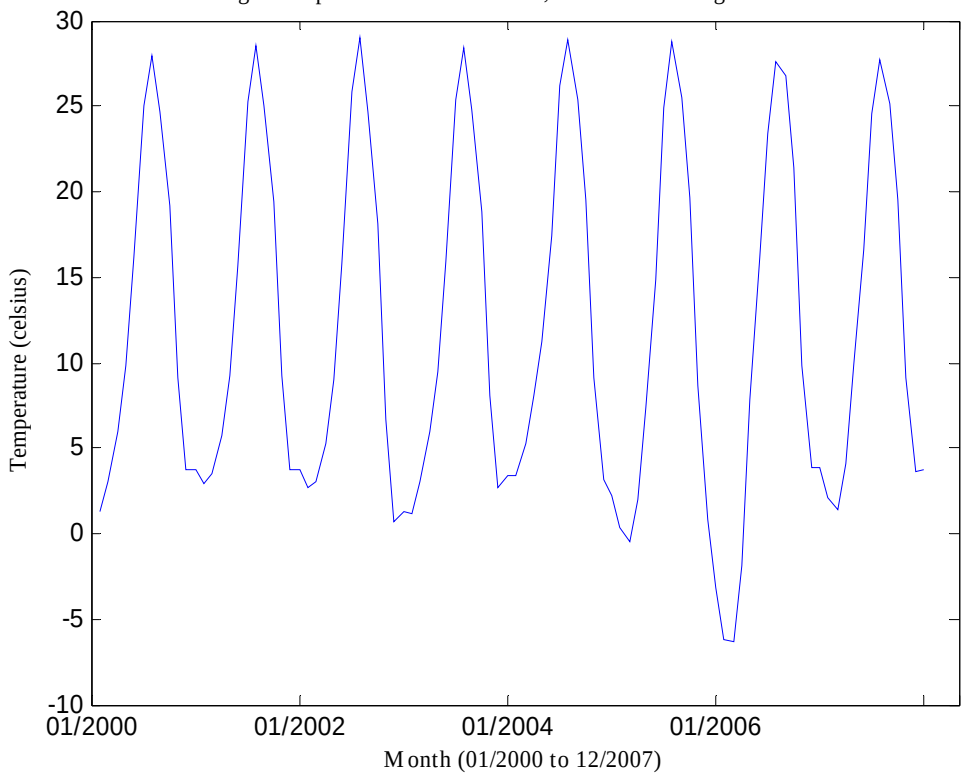

Fig 2. is the monthly average of model CAM-KF of central U.S., from January year 2000 to December year 2007. There are small spikes between January and March, from year 2002 to 2006. These small spikes indicate the temperature of Februarys; therefore, according to the climate model, the monthly average temperature of February is higher than March from year 2002 to 2006. In fact, it is inconsistent with observation. The monthly average of February is lower than March. Fig 3. is the plot of de-noised model output. The small spikes have been smoothed. The de-noised model output suggests monthly average of Februarys is lower than monthly average of March, which is consistent with our observation. It is reasonable to assume that the de-noised model output would give us a better result when we do Bayesian analysis.

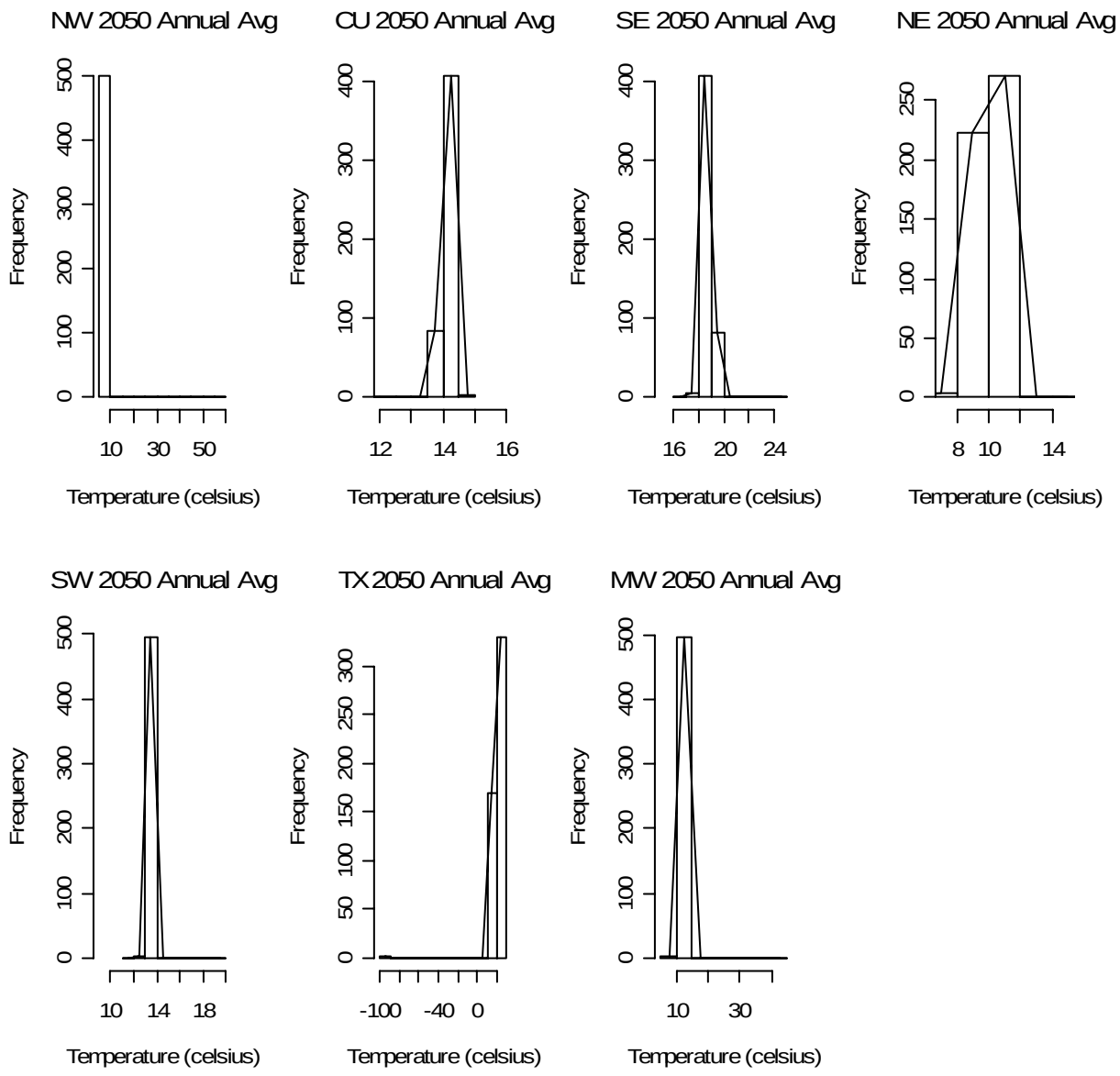

Fig 4. Histograms of projected temperature. NW (Northwest); CU (Central U.S.); SE (Southeast); NE (Northeast); SW (Southwest); TX (Texas); MW (Midwest).

Fig 4 is a plot of histograms of projected annual average temperatures of year 2050. The distributions of the projected temperatures spread a lot, especially in the Northwest, the Southwest, Texas and the Midwest. Fig 6 displays how the data converges. As we have seen from the graph, the projected temperatures spread in most areas. For instance, we find some projected temperatures around -100 degrees in Texas area, which is not acceptable.

After de-noising the model output, we carry out Bayesian analysis. The results are shown in Fig 5 and Fig 7.

In both graphs, the projected temperatures converge better than in Fig 4 and Fig 6. For example the projected temperatures in Texas area are between 19 degrees and 21 degrees; there is no outlier around -100 degrees. The predictions for other areas have also been improved. We find fewer outliers in Fig 5 than in Fig 7.

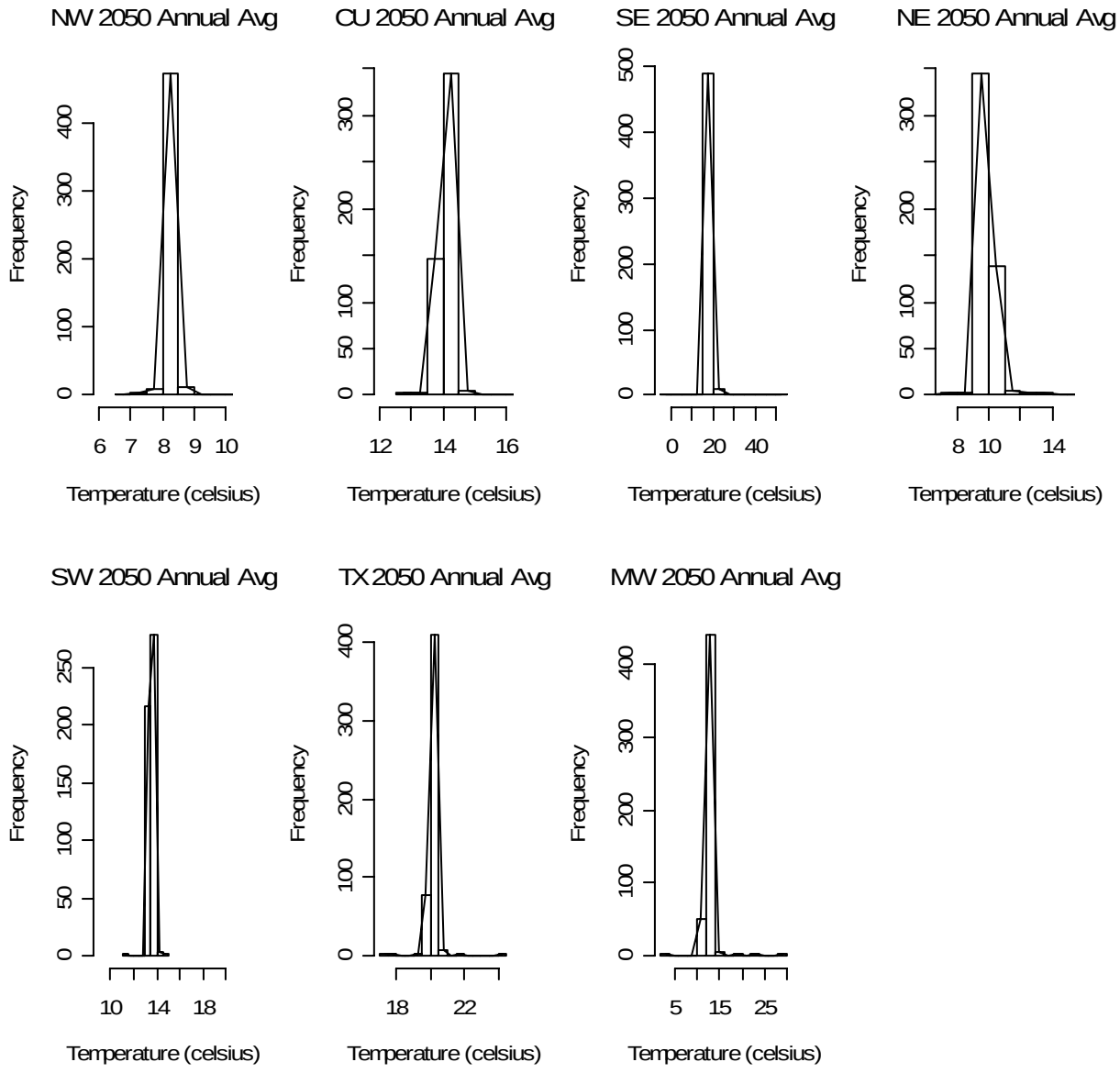

Fig 5. Histograms of projected temperature after the input data being de-noised. NW (Northwest); CU (Central U.S.); SE (Southeast); NE (Northeast); SW (Southwest); TX (Texas); MW (Midwest).

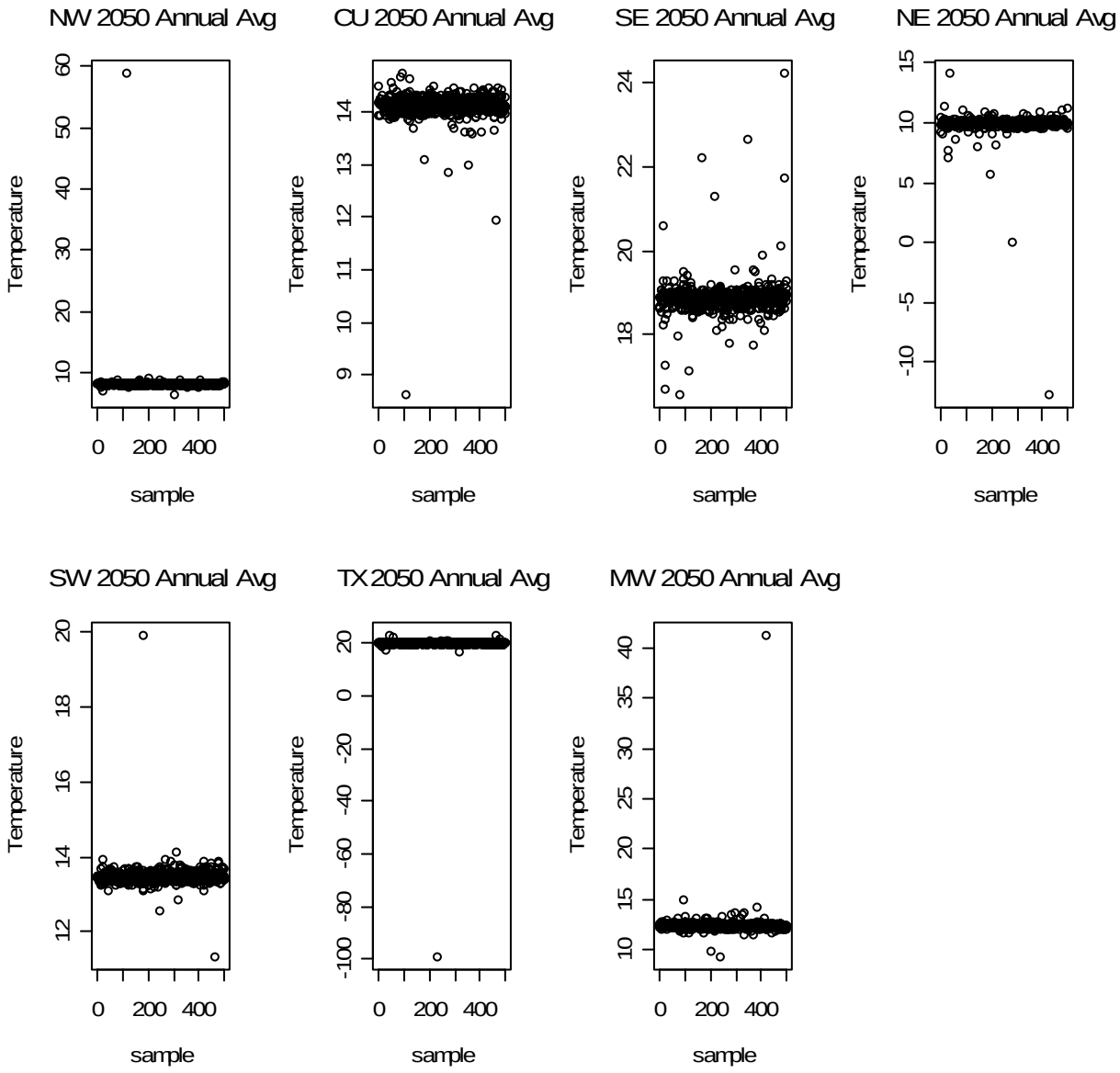

Fig 6. Convergence of projected temperature. NW (Northwest); CU (Central U.S.); SE (Southeast); NE (Northeast); SW (Southwest); TX (Texas); MW (Midwest).

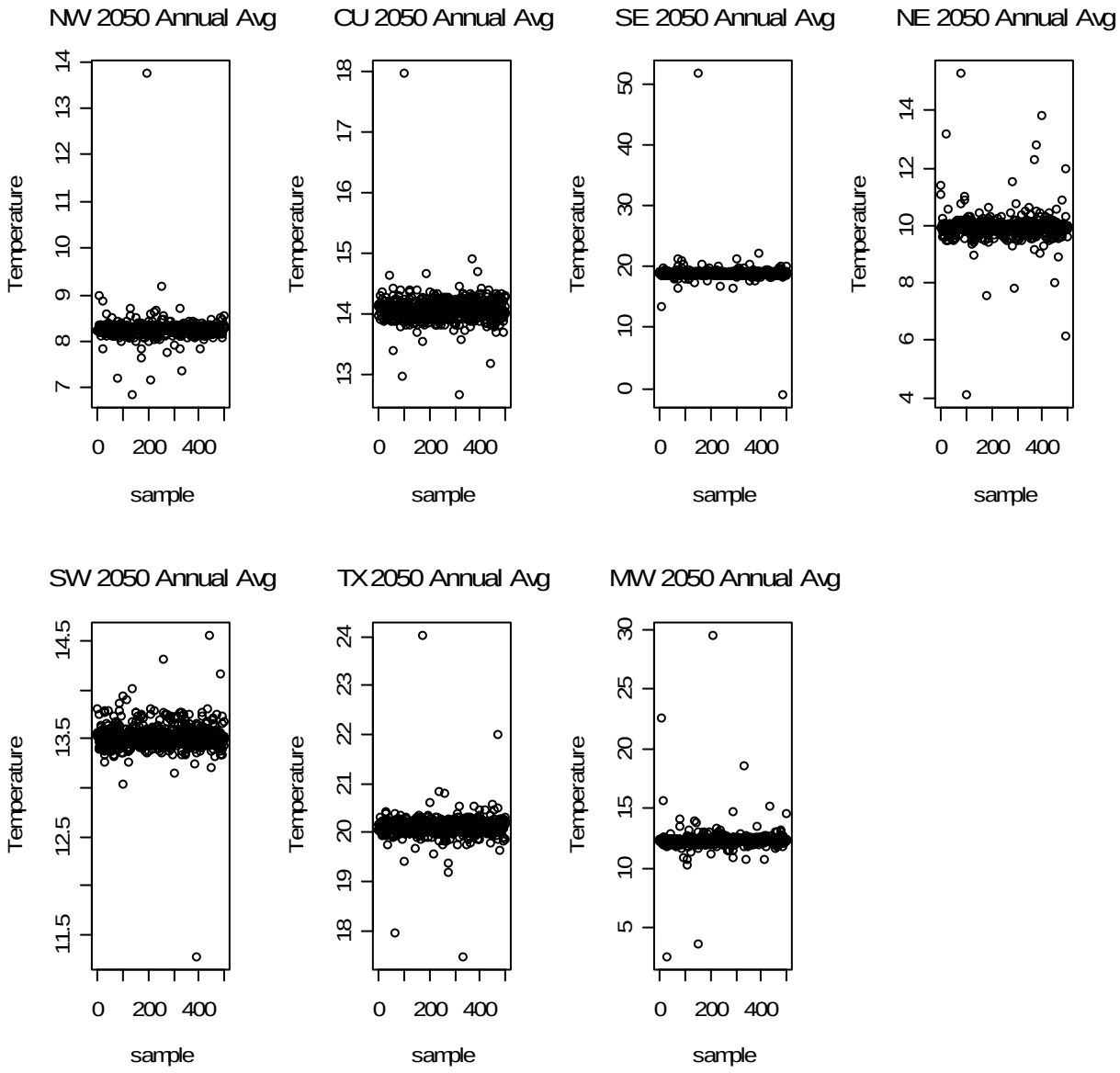

Fig 7. Convergence of projected temperature (data being de-noised before Bayesian analysis). NW (Northwest); CU (Central U.S.); SE (Southeast); NE (Northeast); SW (Southwest); TX (Texas); MW (Midwest).

Discussion and Conclusion

The climate is a dynamical system influenced not only by immense external factors, such as solar radiation, but also by seemingly insignificant phenomena. Even though people want to develop deterministic tools or models to describe the climate system, lots of subtle factors such as butterfly effect, which bring uncertainties to the system, would make the efforts in vain. Many factors, such as background noise and nonlinear components ensure that the climate system is amenable to statistical thinking. Background noise is an internal source of variation in the climate system. Stochastic models are employed in related research. The dynamics of climate are nonlinear. Nonlinear components of the hydrodynamic part include important advective terms. More and more statistical methods have been employed by climate researchers.

We presented a strategy in this paper to the calculation of posterior distributions for future climate change based on an ensemble of WRFs. One feature of our approach is the use of wavelet analysis to de-noise the WRF model output. The output of model CAM_KF, RRTM_KF and RRTM_GRELL for each region (southwest, northwest, texas, central, midwest, southeast, northeast) has been de-noised. Using the de-noised model output, we carry out Bayesian analysis.

Using the de-noised data, MCMC simulation generates a sample of future mean with smaller standard deviation for most regions. Fig 8. compares the convergence of these two different approaches. The de-noised model output converges better in most areas. As for the mean values, no significant difference between these two approaches has been detected. The projected values for future temperature average of these two methods are almost the same.

Table 3. Mean and standard deviation of projected temperature. In most regions, the standard deviation of data being de-noise before adopting Bayesian analysis is smaller than those haven't been de-noised before Bayesian analysis.

	Future mean	future mean (de-noised)	future standard deviation	Future standard deviation (de-noised)
Northwest	8.24954	8.266535	2.271175	0.2928575
Central	14.10099	14.08289	0.3125266	0.2507549
Southeast	18.8731	18.91932	0.4717802	1.786006
Northeast	9.918162	9.934767	1.183863	0.5787154
Southwest	13.47105	13.52534	0.3332825	0.1633585
Texas	19.81693	20.11943	5.334675	0.2890682
Midwest	12.41831	12.37056	1.334683	1.182801

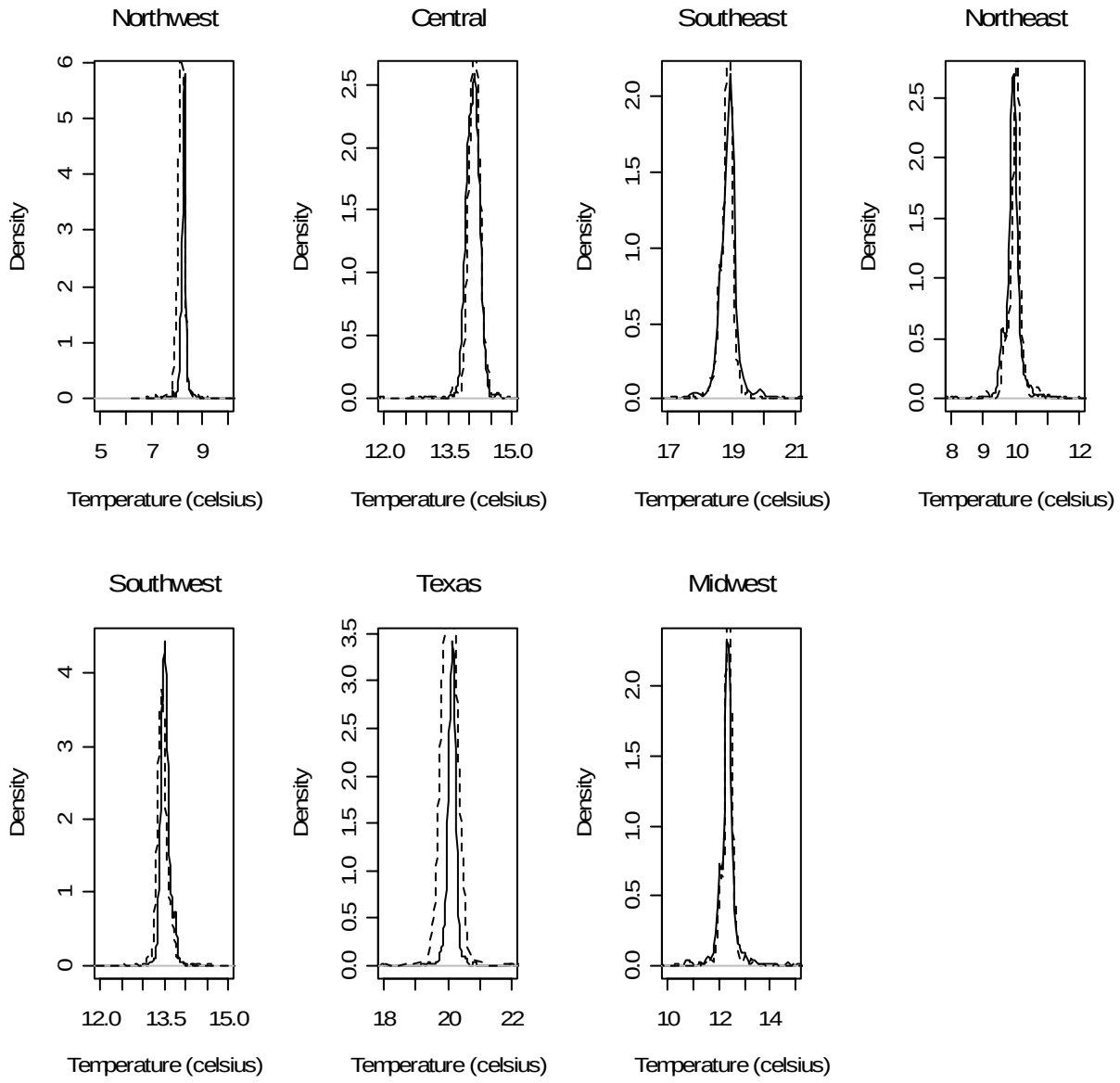

Fig 8. Density plot of annual average of year 2050. The solid lines are the output of Bayesian analysis with input data being de-noised. The dash lines are the output of Bayesian analysis without input data not being de-noised. The solid lines converge better in most cases.

Fig 9. displays our prediction of temperature change from year 2000 to year 2050 in U.S.A. People used to predict the overall future (year 2050) temperature average would be 2.5 to 3 degrees higher than the present (year 2000) average in U.S. Our analysis in this paper gives more details about temperature change. The Midwest area will expect temperature increase by $2.797^{o}c$, which would makes drought severe in the Midwest. Since the interest of this paper is not focusing on climate interpretation, we would just point out the meaning of our research by offering the dataset.

Table 4. Temperature change in U.S. from year 2000 to year 2050.

Region	Temperature Change
Northwest	1.615754
Central	1.88173
Southeast	2.00747
Northeast	2.428731
Southwest	1.69283
Texas	2.12303
Midwest	2.79700

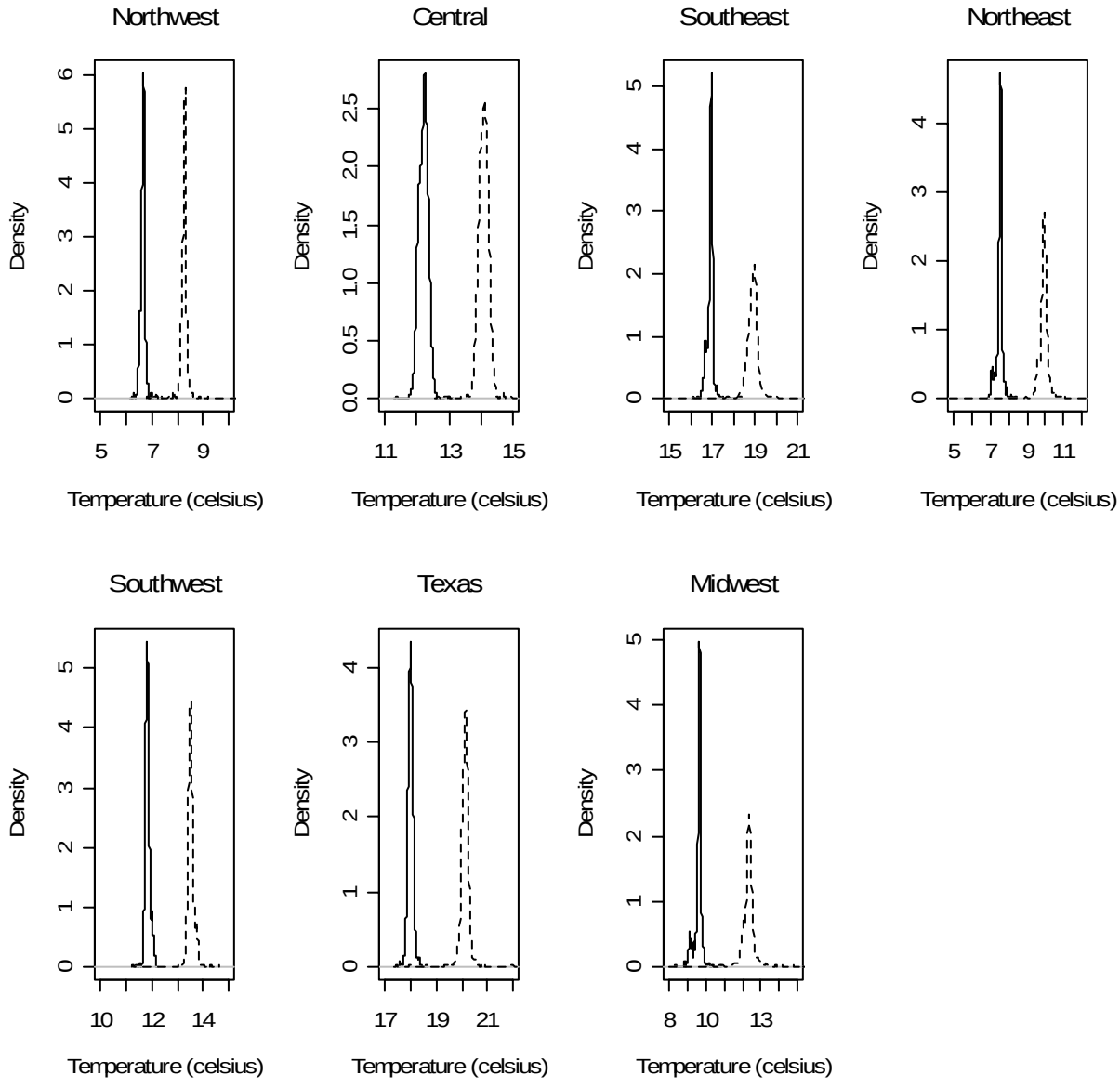

Fig 9. Annual Average Temperature of U.S.A. in year 2000 and 2050. The input data has been de-noised. The solid lines are annual average temperature of year 2000. The dash lines are annual average temperature of year 2050.

Reference

Casella, G., and George, E., (1992): “Explaining the Gibbs Sampler,” The American Statistician, 46, 167-174.

Dessai, S. and Hulme, M. (2004), “Does Climate Adaptation Policy Need Probabilities?” Climate Policy, 4, 107—128.

Donoho, D. L. and Johnstone, I. M. (1992), “New Minimax Theorems, Thresholding, and Adaptation,” Technical Report, Department of Statistics, Stanford Unviersity.

Donoho, D. L. (1995), “De-noising by Soft-thresholding,” IEEE Trans. on Inf. Theory, 41, 3, 613-627.

The Fourth Assessment Report (AR4) of the Intergovernmental Panel on Climate Change.

Gamerman, D., and Lopes, H. F. (2006): “Markov Chain Monte Carlo,” Taylor & Francis Group, .

Gelfand, A.E., and Smith, A.F.M. (1990), “Sampling-based Approaches to Calculating Marginal Densities,” Journal of American Statistical Association, 85, 398-409.

Giorgi, F., and Mearns, L. O. (2002), “Calculation of Average, Uncertainty Range, and Reliability of Regional Climate Changes from AOGCM Simulations via the “reliability ensemble averaging” (REA) Method,” Journal of Climate, 15, 1141-1158.

Green, P. J., (1984), “Iteratively Reweighted Least Squares for Maximum Likelihood Estimation, and Some Robust and Resistant Alternatives”, Journal of Royal Statistical Society, 46B, 149-192.

Smith, R. L., Tebaldi, C., Nychka, D. and Mearns, L. (2009), “Bayesian Modeling of Uncertainty in Ensembles of Climate Models,” Journal of American Statistical Association, 104, 97-116.

Storch, H., and Zwiers, F. W. (2002), “Statistical Analysis in Climate Research,” Cambridge University Press, 2-3.

Tebaldi, C., and Smith, R. L. (2005), “Quantifying uncertainty in projections of regional climate change: a Bayesian approach to the analysis of multimodel ensembles,” Journal of Climate, 18, 1524-1540.

Tebaldi, C., Mearns, L. O., Nychka, D. and Smith, R. L. (2004), “Regional Probability of Precipitation Change: A Bayesian Analysis of Multimodel Simulations,” Geophysics Research Letter, 31, L24213, doi:10.1029/2004GL021276.

The Third Assessment Report (TAR) of the Intergovernmental Panel on Climate Change.

VITA

Yihua Cai was born in Nantong, China on December 20, 1978, the son of Guanming Cai and Haishan Wang. After completing his work at No. 1 Middle School, Nantong, China, in 1997, he entered Nanjing University in Nanjing, China. He received the degree of Bachelor of Arts and Master of Arts from Nanjing University in June, 2001 and June, 2004 respectively. During the following years, he was employed as an editor of English Language at Foreign Language Teaching and Research Press in Beijing, China. In September, 2007, he entered the Graduate School at the University of Texas at Austin.

Permanent Address: Apt. 501
Cheng Gang Xin Cun
Nantong, Jiang Su Province 226001
China

This report was typed by the author.